# GT-Loc: Unifying When and Where in Images Through a Joint Embedding Space

David G. Shatwell[1] Ishan Rajendrakumar Dave[2] Sirnam Swetha[1] Mubarak Shah[1]
[1]Center for Research in Computer Vision, University of Central Florida [2]Adobe
david.shatwell@ucf.edu idave@adobe.com swetha.sirnam@ucf.edu shah@crcv.ucf.edu

## Abstract

*Timestamp prediction aims to determine when an image was captured using only visual information, supporting applications such as metadata correction, retrieval, and digital forensics. In outdoor scenarios, hourly estimates rely on cues like brightness, hue, and shadow positioning, while seasonal changes and weather inform date estimation. However, these visual cues significantly depend on geographic context, closely linking timestamp prediction to geo-localization. To address this interdependence, we introduce GT-Loc, a novel retrieval-based method that jointly predicts the capture time (hour and month) and geo-location (GPS coordinates) of an image. Our approach employs separate encoders for images, time, and location, aligning their embeddings within a shared high-dimensional feature space. Recognizing the cyclical nature of time, instead of conventional contrastive learning with hard positives and negatives, we propose a temporal metric-learning objective providing soft targets by modeling pairwise time differences over a cyclical toroidal surface. We present new benchmarks demonstrating that our joint optimization surpasses previous time prediction methods, even those using the ground-truth geo-location as an input during inference. Additionally, our approach achieves competitive results on standard geo-localization tasks, and the unified embedding space facilitates compositional and text-based image retrieval.*

## 1. Introduction

Estimating the capture time and geo-location of images is crucial for applications ranging from digital forensics to ecological studies and social media management. In digital forensics, accurate timestamps verify image authenticity and help detect manipulation, particularly when camera calibrations are suspect. This capability is essential for reconstructing events from timestamped images during accidents or natural disasters, providing critical information to first responders. Ecological studies benefit from time-ordered images to monitor changes in landscapes and wildlife, while precise timestamps in social media enhance content management and chronological sorting.

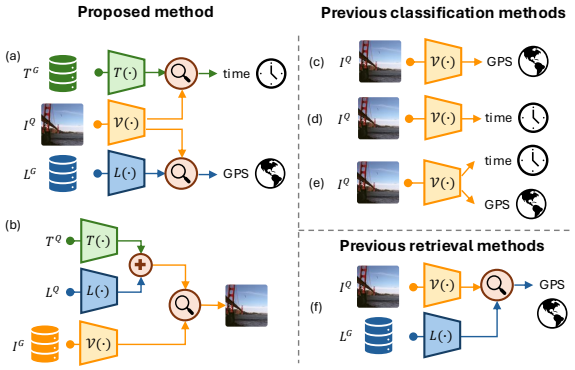


Figure 1. **GT-Loc: Our Unified Approach vs. Prior Methods**. By mapping image, location and time into a single multimodal embedding space, our method can be used for (a) simultaneous image-to-location and image-to-time retrieval, (b) composed geotemporal-to-image retrieval. In contrast, current methods are limited to only (c) location [3, 5], (d) time [27] or (e) geo-temporal classification [45] or (f) image-to-GPS retrieval [11, 37].

Despite its importance, predicting time from images presents several challenges because of the intricate relationship between temporal cues and location-specific factors. Time-of-day (ToD; *i.e.*, hours) and time-of-year (ToY; *i.e.*, months) manifest differently in images due to variables like scene brightness, shadows, weather, and seasonal changes, making it difficult to establish consistent patterns. The complexity of the task is further compounded as the visual appearance of specific hours varies substantially across different months and locations, influenced by the amount and relative exposure to sunlight. Additionally, the representation of months fluctuates across various latitudes, with regions near the equator experiencing relatively stable climate conditions year-round compared to regions at higher latitudes.

Most existing methods [22, 27, 45] rely heavily on GPS metadata for accurate time estimation, while state-of-the-art geo-localization models (e.g., PIGEON [5], GeoCLIP [37]) excel at coarse location prediction. Yet predicting both time and location simultaneously without supplementary inputs remains unsolved, and few works tackle the complementary

task of image retrieval (Figure 1). This challenge is further exacerbated by the absence of standardized datasets and evaluation protocols: many approaches use custom or poorly documented splits, hindering fair comparisons and consistent benchmarking across studies.

In this paper, we introduce *GT-Loc*, a retrieval-based approach for joint time prediction and geo-localization. We conceptualize time prediction as a retrieval problem, representing time as a month-hour pair. A schematic diagram of our framework is shown in Figure 2. Building upon the CLIP-initialized visual model, our goal is to learn a shared embedding space where we can align visual (image), time, and location modalities. For time prediction, we propose a novel time representation that considers the cyclical nature of months and hours over a toroidal manifold. We then project these representation into a multi-scale, high-dimensional time embedding using random Fourier features (RFFs) [33]. Next, to learn the alignment between the time and image embeddings, we explore several possibilities. Existing contrastive learning methods, including CLIP [24] and SimCLR [2], use other batch instances as negative samples. Such a strategy succeeds in image-location losses used by Contrastive Spatial Pre-Training (CSP) [18], GeoCLIP [37], and SatCLIP [11], due to the significant variation of visual appearance with respect to geographical location. In contrast, image-time alignment suffers because temporal neighbors often look nearly identical: hours and months blend smoothly in appearance, so designating adjacent time points as negatives undermines effective alignment. Instead of defining positive-negative pairs as in contrastive learning, we propose a novel *Temporal Metric Learning* approach, which encourages similarity between two instances based on their time difference. To build the target metric for our proposed loss, we use the toroidal distance between the times of each instance pair to consider the cyclic nature of time. This approach enhances performance without the need for explicit assignment of positive and negative samples, providing a more effective and efficient solution for time prediction.

By mapping the image, location, and time modalities into a unified feature space, our model is able to perform compositional retrieval tasks. For instance, given a specific time and location, it can efficiently retrieve all corresponding images from a gallery that closely match the specified criteria.

In summary, our main contributions are the following:

- A framework for joint time-of-capture prediction and geo-localization by aligning the image, time and location embeddings in a shared multimodal feature space using contrastive learning.
- The first retrieval-based method for time-of-capture prediction, where we propose a novel time representation as normalized month-hour pairs, considering its cyclic nature.
- A novel Temporal Metric Learning (TML) loss function for image-time alignment with soft targets, eliminating the need to assign positive and negative samples to the anchor. Since both hours and months are cyclic, we employ a toroidal distance instead of a regular $\ell_2$ distance which results in improved performance.
- A new standard benchmarks for time prediction, demonstrating that our jointly optimized time-location method surpasses time-only optimized baselines and competes well with expert geo-localization methods. Our shared embedding space further facilitates downstream tasks like compositional and text-based retrieval.

## 2. Related Work

**Time-of-capture prediction:** Time-of-capture prediction is a relatively new problem that has only been directly addressed by a handful of prior works. Tsai et al. [36] proposed a physically inspired method to infer the time of day by estimating the Sun's position and camera orientation, but their approach requires sky visibility and additional metadata, such as GPS coordinates and access to an external image database. In a different line of research, Zhai et al. [45] introduced a data-centric approach to learn geo-temporal image features. Their model uses an image, location, and time encoders to generate mid-level features, which are subsequently passed to a set of classifiers for predicting time and location as discrete classes. However, their evaluation shows that providing the location as an input is crucial for predicting the time of day with reasonable accuracy. Similarly, Salem et al. [27] proposed a hierarchical model to predict the month, hour, and week of capture, but this method also assumes known geo-location, limiting its real-world applicability. In contrast, our model relies solely on images to generate accurate time predictions using a retrieval approach in a continuous shared feature space, with resolution determined by a gallery of arbitrary size rather than discrete classes.

Other works have also explored the time-of-capture task indirectly. Li et al. [16] presented an algorithm to verify image capture time and location by comparing the sun position, computed from the claimed time and location, with the actual sun position derived from shadow length and orientation. However, their approach assumes that latitude, time-of-day, and time-of-year are given, with only one potentially corrupted. Padilha et al. [22] proposed a model for time-of-capture verification using a data-centric approach, involving four encoders for ground-level images, timestamps, geo-locations, and satellite images, fed into a binary classifier to predict time consistency. Similarly, Salem et al. [26] proposed a model to generate a global-scale dynamic map of visual appearance by matching visual attributes across images annotated with timestamps and GPS coordinates. Both Padilha et al. [22] and Salem et al. [26] show qualitative

results on time prediction, but are limited by their dependency on geo-location. In contrast, our method accurately estimates both GPS coordinates and timestamps using only images.

Several additional methods explore problems adjacent to time prediction. Jacobs et al. [8] proposed an algorithm for geo-locating static cameras by comparing temporal principal components of yearly image sequences with those from a gallery of known locations. For shadow detection, Lalonde et al. [14] used a multi-stage method to extract pixel-wise ground-shadow features and find edges using CRF optimization, while Wehrwein et al. [40] used illumination ratios to label shadow points in a 3D reconstruction and compute dense shadow labels in pixel space. Another series of works estimate the sun position for various downstream applications, such as computing camera parameters from time-lapses [13] and determining outdoor illumination conditions [6, 14] from single images. Adapting these methods for time prediction would require additional metadata, which might not be available during inference. For example, even with correct sun position prediction, day of the year, geo-location, and compass orientation are needed to accurately predict the hour.

Although the above works contribute to time prediction, they either require additional input metadata like GPS coordinates, or are not reliable in the absence of specific temporal cues. In contrast, our proposed GT-Loc model aims to predict both ToY and ToD from a single image without relying on any additional metadata, making it more broadly applicable for real-time prediction tasks.

**Global geo-localization:** Geo-localization, the task of estimating the geographic coordinates of an image, has gained substantial popularity in recent years. Traditionally, geo-localization methods have adopted either a classification approach [12, 20, 23, 28, 38, 41] or an image retrieval approach [25, 29, 30, 35, 49, 50]. The classification approach divides the Earth into a fixed number of geo-cells or uses the city label, assigning the center coordinate of the selected class as the GPS prediction. However, this can result in significant errors depending on the size of the geo-cells, even when the correct class is selected. In contrast, the image retrieval approach compares a query image to a gallery and retrieves the image with the highest similarity. GeoCLIP [37] addresses the limitations of traditional approaches by framing global geo-localization as a GPS retrieval problem. It leverages the pretrained CLIP [24] ViT and employs contrastive learning to align image and location embeddings in a shared feature space. Other methods, such as PIGEON [5], use a hybrid strategy: first using image classification to identify the top-$k$ geo-cells with the highest probability, followed by a secondary retrieval stage for refinement within and across geo-cells. However, PIGEON's dependence on additional metadata for training—such as administrative boundaries, climate, and traffic—poses a significant limitation. Finally, recent methods such as Img2Loc [47] have begun leveraging multimodal large language models (MLLMs) and retrieval-augmented generation (RAG) to achieve competitive geo-localization performance without the need for dedicated training. However, the effectiveness of these methods is highly dependent on the underlying MLLM and results in substantial inference overhead.

**Geo-spatial dual-encoder methods:** The success of CLIP has inspired to leverage its architecture for geo-spatial tasks. SatCLIP [11] aligns satellite imagery and natural images in a shared feature space, enabling cross-modal retrieval and localization. In a similar fashion, Zavras et al. [43] proposed a method for aligning complementary remote sensing modalities beyond RGB with the CLIP encoders. Other works from Mai et al. [18] and Mac Aodha et al. [17], employ a dual image-location encoder architecture to learn robust location representations from images. However, their ultimate goal is not to geo-locate images. Instead, they use the learned embeddings from the image encoder for downstream tasks, such as image classification. These methods demonstrate the effectiveness of dual-encoder architectures for geo-spatial problems, motivating our approach of using a triple encoder architecture for joint time-of-capture and location prediction.

## 3. Method

Given a training dataset $S_{train} = \{(I_i, G_i, D_i)\}_{i=1}^{N}$ consisting of image $I_i$, GPS coordinates $G_i$ and date-time $D_i$ triplets, our objective is to train a model that can simultaneously predict the location, time-of-day (ToD) and time-of-year (ToY) from unseen images. Our GT-Loc method consists of three encoders: Image Encoder ($\mathcal{V}$), Location Encoder ($\mathscr{L}$), and Time Encoder ($\mathcal{T}$) as shown in Figure 2. Both geo-localization and time prediction are framed as a retrieval problem. Given a query image $I^Q \in S_{eval}$, we compute an image embedding, $V^Q = \mathcal{V}(I^Q)$, using a pre-trained Vision Transformer. Similarly, given a gallery of latitude-longitude pairs, and a gallery of timestamps, we respectively compute galleries of location embeddings $L_k^Q = \mathscr{L}(G_k^Q)$ and time embeddings $T_k^Q = \mathcal{T}(D_k^Q)$. In order to predict the location and time, the image embedding is compared against both galleries. The GPS and timestamp with the highest cosine similarity to the image are selected as predictions.

A fundamental prerequisite for retrieval is the alignment of image, location, and time modalities within a shared multimodal embedding space. To achieve this, our framework is optimized using two multimodal alignment objectives: (1) Image-Location alignment, and (2) Image-Time alignment. Furthermore, to capitalize on large-scale visual pretraining, we employ the pretrained CLIP ViT-L/14 as our image encoder, projecting it into our shared embedding space using a trainable multi-layer perceptron (MLP).

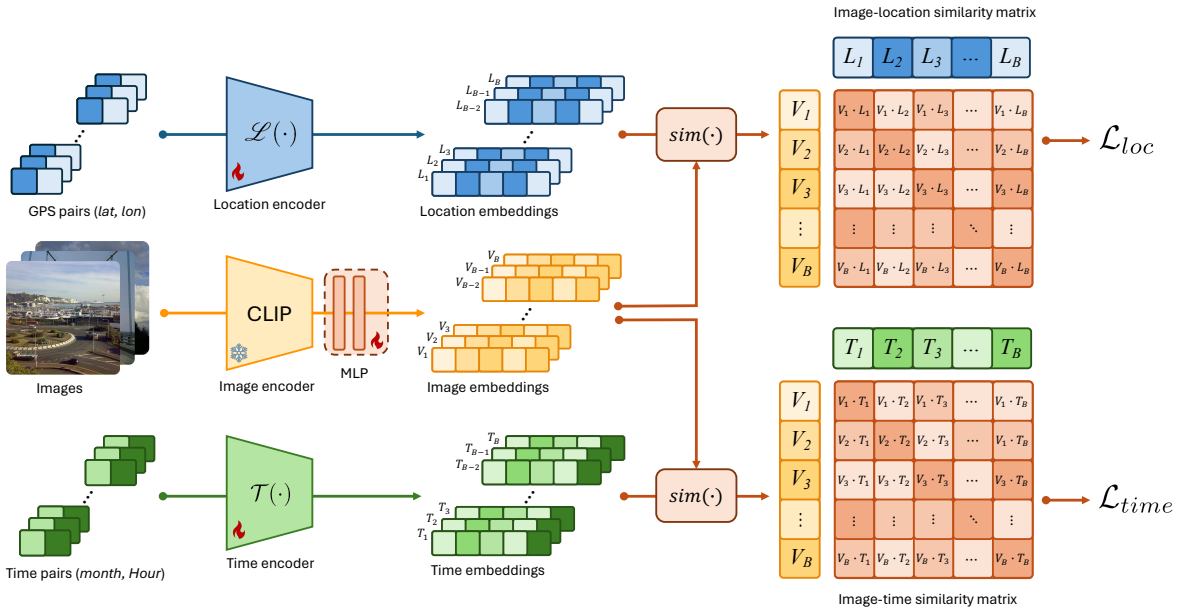


Figure 2. **Overview of GT-Loc**: GT-Loc uses an image encoder $\mathcal{V}(\cdot)$, location encoder $\mathscr{L}(\cdot)$ and time encoder $\mathcal{T}(\cdot)$ to generate a set of image $V_i$, location $L_i$ and time $T_i$ embeddings. Leveraging the CLIP [24] pretrained ViT-L/14 as image encoder, we aim to align its image embedding to both location and time embeddings. The image-location alignment is learned through a regular CLIP-like loss [37] and the image-time alignment is learned through our proposed *Temporal Metric Learning*.

### 3.1. Image-Location Alignment

We adopt GeoCLIP for the image-location modality alignment. Given a latitude-longitude pair $G_i$, it first uses Equal Earth Projection (EEP) to mitigate the distortion of the standard GPS coordinate system and provide a more accurate representation $G_i'$. Then, Random Fourier Features (RFF) are used to map the 2D representation into a rich high-dimensional representation at three scales ($M$) using projection matrices $\gamma(\cdot)$ with different frequencies $\sigma_i \in \{2^0, 2^4, 2^8\}$. Lastly, the RFFs are passed to a set of MLPs $f_i$ and added together, forming a single multi-scale feature vector. This can be mathematically expressed as the following equation:

$$L_i = \mathscr{L}(G_i) = \sum_{i=1}^{M} f_i(\gamma(EEP(G_i), \sigma_i)). \qquad (1)$$

Next, to compute the image-location contrastive loss $\mathcal{L}_{loc}^{i}$, we consider a set of $P$ augmented image $V_{ij}$ and location $L_{ij}$ embeddings ($j \in 1, \ldots, P$). For the batch with size $B$ with $S$ additional location embeddings stored in a continually updated dynamic queue, and temperature $\tau$, the loss is given by:

$$\mathcal{L}_{loc}^{(i)} = -\sum_{j=1}^{P} \log \left( \frac{\exp(V_{ij} \cdot L_{ij}/\tau)}{\sum_{k=1}^{B+S} \exp(V_{kj} \cdot L_{kj}/\tau)} \right). \qquad (2)$$

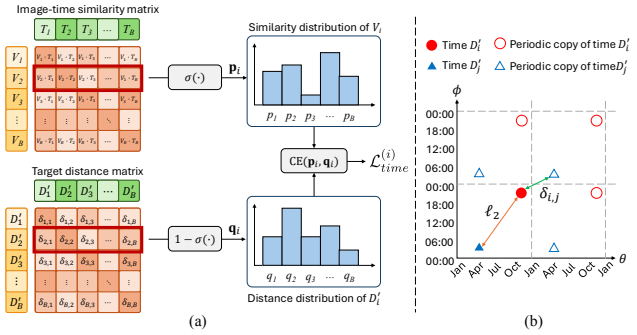


Figure 3. (a) **Temporal Metric Learning loss $\mathcal{L}_{time}$:** We compute the **image-time similarity matrix** by using the cosine similarity between the image and time embedding of the all instances in the batch. We then obtain the **target distance matrix** by computing the cyclic toroidal time difference between each pair. As shown in the red highlighted box, we take the $i$th row of both matrices and normalize them using the softmax function $\sigma(\cdot)$ and $1-\sigma(\cdot)$ respectively, resulting in two probability mass functions $\mathbf{p}_i$ and $\mathbf{q}_i$. The loss is defined as the cross-entropy (CE) between $\mathbf{p}_i$ and $\mathbf{q}_i$. **(b) Proposed distance metric** $\delta_{i,j}$: Assume we have two normalized month-hour pairs $D_i' = (\theta_i, \phi_i)$ and $D_j' = (\theta_j, \phi_j)$. Since month and hours are periodic, they repeat infinitely in both dimensions of the $\theta$-$\phi$ plane. Using the $\ell_2$ distance overestimates the real distance between the two times, but our proposed toroidal time distance $\delta_{i,j}$ is able to provide a correct estimate by considering the minimum distance between $D_i'$ and the periodic copies of $D_j'$.

### 3.2. Time Representation

The capture time of an image is usually a Unix timestamp, an integer tracking the seconds (or milliseconds) elapsed since January 1, 1970. We discard the year information from the timestamp, as predicting the year is beyond the scope of this work, and instead focus on Time-of-Year (ToY; *i.e.*, month) and Time-of-Day prediction (ToD; *i.e.*, hour). Both ToY and ToD are cyclical, with periods of 12 months and 24 hours, respectively. To convert the Unix timestamp $U_i$ into a time representation that focuses on the months and hours, we transform it into a date-time tuple $D_i = \texttt{unix2tuple}(U_i) = (m_i, d_i, H_i, M_i, S_i)$ with the month, day, hour, minute, and second. From this tuple, we then compute a new time representation composed of the normalized cyclic month-hour pair $D'_i = (\theta_i, \phi_i)$ using

$$\theta_i = \frac{1}{12}\left((m_i - 1) + \frac{(d_i - 1)}{\mathcal{D}(m_i)}\right), \tag{3}$$

$$\phi_i = \frac{1}{24}\left(H_i + \frac{M_i}{60} + \frac{S_i}{3600}\right), \tag{4}$$

where $\mathcal{D}(m_i)$ is the number of days in month $m_i$. We represent the sequence of operations to convert a Unix timestamp to normalized cyclic month-hour pair as $D'_i = \texttt{unix2cyclic}(U_i)$.

**Time encoding:** By representing time as a pair of real-valued numbers, the problem of time prediction becomes similar in nature to geo-localization. Instead of retrieving the latitude-longitude pair with the highest similarity, we are interested in retrieving a month-hour pair. Thus, our time encoder ($\mathcal{T}$) follows the exact same architecture as the location encoder ($\mathscr{L}$). Similar to Eq. 1, the time embedding is obtained from the proposed time representation using the following equation:

$$T_i = \mathcal{T}(U_i) = \sum_{i=1}^{M} f_i(\gamma(\texttt{unix2cyclic}(U_i), \sigma_i)). \tag{5}$$

### 3.3. Temporal Metric Learning

Visual features associated with location typically exhibit significant variation due to cultural, socio-economic and environmental factors, leading to abrupt changes in visual embeddings with respect to spatial distances. For instance, two neighborhoods within the same city often appear notably different. Therefore, standard contrastive objectives that clearly distinguish positives and negatives are suitable for aligning location and image embeddings (Eq. 2). In contrast, visual cues change more smoothly and continuously over time. As a result, labeling temporally adjacent samples as negatives, as is common in standard contrastive losses, can hinder effective learning. This makes it challenging to define explicit positives and negatives for time-image alignment. To address this, we propose a metric-learning objective called *Temporal Metric Learning* (TML), explicitly designed to leverage the smooth and cyclic nature of time by aligning instance similarity proportionally to the temporal difference.

Let's consider the image embeddings $\{V_i\}_{i=1}^{B}$ and time embeddings $\{T_i\}_{i=1}^{B}$. Instead of defining a set of positive and negative pairs for each embedding, we assign soft targets inversely proportional to the difference between the time associated to the image and time embeddings. Since months and hours are cyclical, using the regular $\ell_2$ distance between two normalized month-hour pairs $(\theta_i, \phi_i)$ and $(\theta_j, \phi_j)$ in an Euclidean space results in overestimated distance values, as shown in Figure 3(b). This problem can be solved by mapping the normalized month-hour pairs into the surface of a toroidal manifold, resulting in the new distance $\delta_{i,j}$:

$$\delta_{i,j} = \sqrt{\sum_{\alpha \in \{\theta, \phi\}} \min(1 - |\Delta\alpha_{i,j}|, |\Delta\alpha_{i,j}|)^2}. \tag{6}$$

where $\Delta\alpha_{i,j} = \alpha_i - \alpha_j$. Then, for each anchor image embedding $V_i$, we compute a vector $\mathbf{p}_i$ with the normalized cosine similarity scores with respect to all time embeddings $T_j$ in the batch. Similarly, we compute the vector $\mathbf{q}_i$ with the normalized time difference between the anchor time and other times in the batch as follows:

$$\mathbf{p}_i[j] = \frac{\exp\left(V_i \cdot T_j / \tau\right)}{\sum_{k=1}^{B} \exp\left(V_i \cdot T_k / \tau\right)}, \quad j \in [B], \tag{7}$$

$$\mathbf{q}_i[j] = 1 - \frac{\exp\left(\delta_{i,j}\right)}{\sum_{k=1}^{B} \exp\left(\delta_{i,k}\right)}, \quad j \in [B]. \tag{8}$$

Since $\mathbf{p}_i$ and $\mathbf{q}_i$ are normalized, they have the same characteristics as probability mass functions. Thus, we define the image-time contrastive loss by computing the cross-entropy (CE) between $\mathbf{p}_i$ and $\mathbf{q}_i$:

$$\mathcal{L}_{time}^{(i)} = \text{CE}(\mathbf{p}_i, \mathbf{q}_i). \tag{9}$$

The final training objective is defined as the sum of the image-location and image-time objectives, defined in Equations 2 and 10:

$$\mathcal{L}^{(i)} = \mathcal{L}_{loc}^{(i)} + \mathcal{L}_{time}^{(i)}. \tag{10}$$

### 3.4. Inference

After training, the image, location, and time modalities share a unified embedding space. To predict the capture time and location of a query image $I^Q$, we compute the cosine similarity between its image embedding $V^Q$ and the embeddings within the time gallery $T^G$ and the location gallery $L^G$. The predicted time and location correspond to the gallery elements with the highest cosine similarities. For a visual representation of this inference process, refer to Supplementary Section 7.

Table 1. **Zero-shot time prediction** on the *unseen* cameras of SkyFinder dataset. Rows marked by * indicate methods we replicate, closely adhering to the protocols outlined by prior work.

| Method | Month Error ↓ | Hour Error ↓ | TPS ↑ |
|---|---|---|---|
| Zhai et al. [45]* | 2.46 | 3.18 | 65.48 |
| Padilha et al. [22]* | 1.62 | 3.61 | 71.42 |
| Salem et al. [27]* | 2.72 | 3.05 | 63.25 |
| Zhai et al. [45]* w/ CLIP | 1.65 | 3.14 | 73.20 |
| Padilha et al. [22]* w/ CLIP | 1.62 | 2.98 | 74.06 |
| Salem et al. [27]* w/ CLIP | 1.56 | 2.87 | 75.02 |
| CLIP + cls. head | 1.61 | 3.17 | 73.37 |
| CLIP + reg. head | 1.55 | 3.06 | 74.33 |
| DINOv2 + cls. head | 2.24 | 3.74 | 65.61 |
| DINOv2 + reg. head | 2.12 | 3.53 | 67.49 |
| OpenCLIP + cls. head | 1.66 | 3.43 | 71.87 |
| OpenCLIP + reg. head | 1.72 | 3.34 | 71.75 |
| TimeLoc | 1.52 | 2.84 | 75.49 |
| **GT-Loc (Ours)** | **1.40** | **2.72** | **77.00** |

Table 2. **Geo-localization accuracy** on Im2GPS3k & GWS15k datasets, reported on the ratio of samples that are correctly predicted under a distance threshold of 1 km radius.

| Method | Im2GPS3k | GWS15k |
|---|---|---|
| [L] kNN, sigma=4 ICCV'17[38] | 7.2 | - |
| PlaNet ECCV'16[41] | 8.5 | - |
| CPlaNet ECCV'18[28] | 10.2 | - |
| ISNs ECCV'18[20] | 10.5 | 0.1 |
| Translocator ECCV'22[23] | 11.8 | 0.5 |
| GeoDecoder CVPR'23[3] | 12.8 | 0.7 |
| GeoCLIP NeuRIPS'24[37] | 14.11 | 0.6 |
| PIGEOTTO CVPR'24[5] | 11.3 | 0.7 |
| Img2Loc(LLaVA) SIGIR'24[47] | 8.0 | - |
| **GT-Loc (Ours)** | **14.41** | **0.88** |

## 4. Experiments

We evaluate GT-Loc on both time-of-capture and geo-localization tasks, conduct ablation studies to justify key design choices, and assess robustness by measuring performance under limited training data and noisy annotations.

**Datasets and evaluation details**: For training, we use two existing datasets: MediaEval Placing Tasks 2016 (MP-16) [15] and Cross-View Time (CVT) [26]. MP-16 consists of 4.72 Million images from Flickr annotated only with GPS coordinates. CVT originally consists of 206k geo-tagged smartphone pictures from the Yahoo Flickr Creative Commons 100 Million Dataset [34] and 98k images from static outdoor webcams of the SkyFinder Dataset [19]. Our zero-shot evaluation benchmark is constructed using a subset of images from SkyFinder that are not seen during training.

Geo-localization performance is evaluated by measuring the geodesic distance between the real and predicted GPS coordinates, and then computing the ratio of images that are correctly predicted within a threshold. Time prediction performance is evaluated by measuring the mean absolute ToY ($E_{ToY}$) and ToD ($E_{ToD}$) errors between the ground-truth and predicted times. We also report an overall **Time Prediction Score (TPS)** that combines both errors into a single metric. We compute the TPS based on our proposed cyclical time difference using the following equation:

$$TPS = 1 - \sqrt{\frac{\widetilde{E}_{ToY}^2 + \widetilde{E}_{ToD}^2}{2}}, \tag{11}$$

where $\widetilde{E}_{ToY}, \widetilde{E}_{ToD} \in [0, 1]$ are the normalized time errors. A TPS of 1 represents a perfect prediction, while 0 represents the maximum possible cyclic error of 12 hours and 6 months. Please, refer to Supplementary Sections 8 to 12 for more details about the dataset, architecture and training protocol.

### 4.1. Comparison with Prior Methods

**Time-of-capture prediction:** We present our time prediction results in Table 1. Given the significant reproducibility challenges identified in prior work (see Supplementary Section 21), we selected three representative baseline methods for a fair comparison with GT-Loc. The first baseline is the triple encoder architecture from Zhai et al. [45], which closely resembles our approach. Although they did not release code, pretrained weights, or exact dataset splits, their methodology is well-documented, allowing us to closely replicate their protocol. The second baseline is Padilha et al. [22], chosen due to its recent publication, public source code availability, and use of the CVT dataset—common to our evaluation. However, they do not provide the specific cross-camera split of CVT, and their original work only offers qualitative results. We extend their evaluation quantitatively for direct comparison. The final baseline is Salem et al. [27], selected for its use of the SkyFinder dataset, with clear experimental procedures enabling straightforward replication despite the absence of source code. Importantly, all these baseline methods rely on geo-location metadata as input, unlike GT-Loc, which exclusively uses visual cues. Additionally, each baseline fully trains their models using different backbone architectures: Zhai et al. [45] employs InceptionV2, while both Padilha et al. [22] and Salem et al. [27] use DenseNet-121. To ensure fairness, we also evaluate scenarios in which these models use a frozen CLIP ViT-L/14 backbone, aligning closely with GT-Loc.

Our second set of baselines explores various frozen backbone architectures (e.g., CLIP ViT-L/14, DINOv2 ViT-L/14, and OpenCLIP ViT-G/14), paired with an MLP of matching capacity to our image encoder, and either regression or

classification heads. These baselines illustrate the effectiveness of our retrieval-based approach compared to standard classification and regression techniques.

Lastly, we introduce TimeLoc, a robust baseline utilizing only the image and time encoders from GT-Loc, analogous to GeoCLIP [37] and SatCLIP [11] but specifically tailored for time prediction. Experimental results clearly demonstrate that GT-Loc achieves superior accuracy for both Time-of-Year (ToY) and Time-of-Day (ToD) predictions without additional metadata. Moreover, jointly training for time prediction and geo-localization enriches the learned temporal representations. In Figure 4 , we show two qualitative examples from unseen cameras of the SkyFinder dataset, highlighting GT-Loc's effectiveness in simultaneously predicting time and geo-location.

Table 3. **Ablations** for time prediction performance with different loss functions. $\ell_2$ refers to the Euclidean distance, while cyclic refers to the distance over the toroidal manifold.

| Loss Function | Month Error ↓ | Hour Error ↑ | TPS ↓ |
|---|---|---|---|
| CLIP [24] | 1.71 | 3.51 | 71.12 |
| RnC [44] | 1.87 | 2.96 | 71.89 |
| SimCLR [2] | 1.50 | 3.40 | 73.28 |
| TML ($\ell_2$) | 1.53 | 2.74 | 75.88 |
| **TML (Cyclic)** | **1.40** | **2.72** | **77.00** |

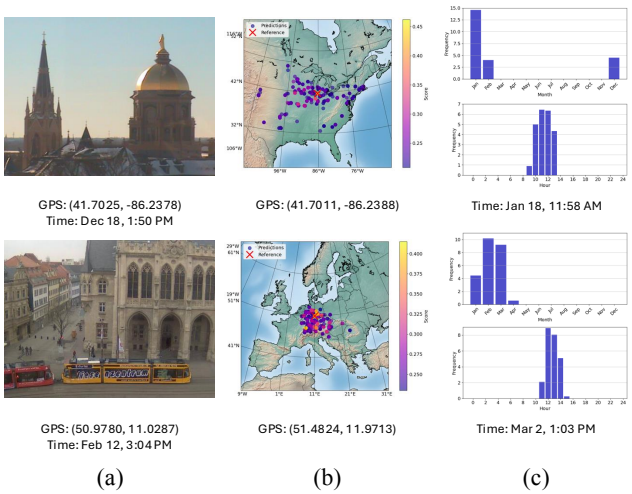


Figure 4. (a) Sample images for two cameras of the SkyFinder test set with the ground truth location and capture time. (b) Spatial distribution of the predicted GPS coordinates colored by the cosine similarity between the location and image embeddings. (c) Histogram of the top-1k retrieved months and hours, weighted by the cosine similarity between the image and time embeddings (supp. Eq. 12). Both the top-1 predicted location and time are shown below the distributions.

**Geo-localization:** Table 2 presents the geo-localization performance of our model compared to recent expert methods. Overall, GT-Loc delivers competitive results compared against models like GeoCLIP on both datasets. It also outperforms the LLaVA-based Img2Loc variant, despite using at least 15 times fewer parameters.

## 4.2. Ablation Studies

In this section, we present additional experiments to evaluate the effectiveness of our proposed Temporal Metric Loss. Furthermore, in the supplementary material (Section 13), we explore alternative image backbones (*i.e.*, DINOv2 [21], OpenCLIP [7]), different time encoders (*i.e.*, Time2Vec [10], Circular Decomposition [17]), and various temporal resolutions to provide additional insights and justification for our design choices.

We evaluate several alternative loss functions for time prediction. First, we implement a basic CLIP-based loss [24]. Next, we explore a geo-localization-style contrastive loss that uses a dynamic queue and applies a false negative mask, excluding samples close to the anchor based on a specified threshold. We also experiment with the Rank-N-Contrast loss introduced by Zha et al. [44], specifically designed for regression tasks by ranking samples according to their distances. Finally, we compare our proposed loss function against a variant using standard $\ell_2$ distance rather than cyclic temporal distance. The results in Table 3 demonstrate that our Temporal Metric Loss significantly outperforms all other evaluated losses for both Time-of-Day (ToD) and Time-of-Year (ToY) predictions.

Table 4. **Impact of Limited Data** on the Robustness of Time Prediction

| Data Availability | Month Error ↓ | Hour Error ↓ | TPS ↑ |
|---|---|---|---|
| 100% | 1.40 | 2.72 | 77.00 |
| 50% | 1.69 | 2.83 | 74.02 |
| 10% | 1.70 | 2.94 | 73.51 |
| 5% | 1.89 | 2.86 | 72.07 |

Table 5. **Ablations** for robustness to label noise for time-prediction.

| Label Noise ($\sigma$) | Month Error ↓ | Hour Error ↓ | TPS ↑ |
|---|---|---|---|
| 0 | 1.40 | 2.72 | 77.00 |
| 1 | 1.52 | 2.71 | 76.00 |
| 2 | 1.75 | 2.74 | 73.81 |
| 3 | 2.16 | 2.72 | 69.92 |

## 4.3. Robustness of GT-Loc

Timestamp-prediction datasets are often scarce or plagued by missing and noisy metadata, so robustness is essential.

We first measure performance as we shrink the training set from 100% to 5% in four stages, then simulate label noise by adding Gaussian perturbations with standard deviation between $\sigma \in [0, 3]$ months and hours to the training annotations. As Tables 4 and 5 show, GT-Loc's errors rise only modestly by 0.3 months and 0.22 hours, even with just 5% of the data, and it remains stable up to $\sigma = 2$, with significant degradation appearing only at $\sigma \geq 3$. These results confirm GT-Loc's resilience under realistic data constraints.

### 4.4. Compositional Image Retrieval

In this section, we evaluate GT-Loc's capability for compositional image retrieval tasks, leveraging its unified multimodal embedding space. Specifically, we explore retrieving images based on joint queries consisting of both location and time information. Given query location $L^Q$ and time $T^Q$, we generate a multimodal representation by averaging their embeddings, inspired by multimodal retrieval methods proposed in prior works [31, 32].

To provide meaningful baselines for comparison, we selected the method by Zhai et al. [45], which is conceptually similar but was originally intended only for time and location classification tasks. Since their original work did not consider compositional retrieval explicitly, we adapt their model to our retrieval scenario as described in Supplementary section 14. We evaluate retrieval performance using recall metrics at ranks 1, 5, and 10. A retrieved image is considered correct if its ground truth timestamp is within one hour and one month, respectively, of the query, and if its location is within 25 km.

Table 1 shows quantitative results for these baselines. GT-Loc consistently achieves higher recall across all ranks, significantly outperforming other methods. This indicates that our unified retrieval-based embedding approach effectively encodes joint geo-temporal information, providing richer representations for retrieval tasks. Qualitative examples demonstrating GT-Loc's compositional retrieval capabilities are shown in Supplementary section 14.

Table 6. **Zero-shot composed retrieval** ($T + L \rightarrow I$) on the unseen cameras of the SkyFinder dataset.

| Method | R@1 | R@5 | R@10 |
|---|---|---|---|
| Zhai et al. [45]* | 0.91 | 7.81 | 13.61 |
| Zhai et al. [45]* w/ CLIP | 2.58 | 16.22 | 29.46 |
| **GT-Loc** | **6.69** | **24.58** | **38.54** |

### 4.5. Qualitative results using text queries

We also investigate the ability of GT-Loc to use the pretrained CLIP text encoder to retrieve times and locations mentioned in the text. For this task, we follow GeoCLIP's approach and replace the image backbone by a text backbone, keeping the trained MLP, location encoder and time encoder. For each text, we create a text embedding, pass it through the MLP and compare it against the location and time galleries. We then create spatial and temporal distributions of the top retrieved samples for each modality, as shown Figure 5, where we see that not only is our model able to accurately pinpoint the location, but it also creates meaningful time distributions to words such as "winter" and "evening" that do not explicitly mention the time.

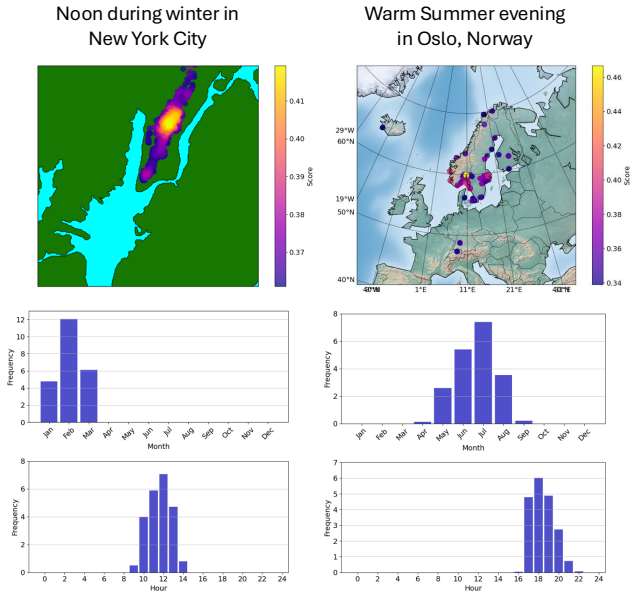


Figure 5. Qualitative examples of geo-localization and time-of-capture prediction using text queries. Top: prompt passed to CLIP's text encoder. Middle: spatial distribution of the predicted geolocations with the highest cosine similarity. Bottom: histogram of the top-1k predicted months and hours with the highest cosine similarity. GT-Loc is capable of providing good estimates of the time and GPS coordinates to each of the text queries, even when time is not explicitly specified.

## 5. Conclusion

We introduce GT-Loc, a novel framework for jointly predicting the time and location of an image using a retrieval approach. GT-Loc not only shows competitive performance compared to state-of-the-art geo-localization models but also introduces the capability of precise time-of-capture predictions. A key innovation of our approach is the novel temporal metric loss, which significantly outperforms traditional contrastive losses in time prediction tasks.

Furthermore, our results demonstrate that GT-Loc extends beyond standard time-of-capture prediction and geo-localization tasks. It supports additional functionalities like compositional image retrieval ($T + L \rightarrow I$), as well as text-to-location and text-to-image retrieval, indicating a profound understanding of the interplay between images, locations, and time.

## 6. Acknowledgements

Supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

## References

[1] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, and Loic Landrieu. OpenStreetView-5M: The many roads to global visual geolocation. *CVPR*, 2024. 4, 5

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 7

[3] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23182–23190, 2023. 1, 6

[4] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 5

[5] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12893–12902, 2024. 1, 3, 6

[6] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7312–7321, 2017. 3, 7

[7] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 7, 2

[8] Nathan Jacobs, Scott Satkin, Nathaniel Roman, Richard Speyer, and Robert Pless. Geolocating static cameras. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–6. IEEE, 2007. 3, 6

[9] Nathan Jacobs, Walker Burgin, Nick Fridrich, Austin Abrams, Kylia Miskell, Bobby H. Braswell, Andrew D. Richardson, and Robert Pless. The global network of outdoor webcams: Properties and applications. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, pages 111–120, 2009. 6

[10] Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019. 7, 2

[11] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023. 1, 2, 3, 7, 5

[12] Parth Parag Kulkarni, Gaurav Kumar Nayak, and Mubarak Shah. Cityguessr: City-level video geo-localization on a global scale. In *European Conference on Computer Vision*, pages 293–311. Springer, 2024. 3

[13] Jean-François Lalonde, Srinivasa G Narasimhan, and Alexei A Efros. What do the sun and the sky tell us about the camera? *International Journal of Computer Vision*, 88: 24–51, 2010. 3

[14] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision*, 98:123–145, 2012. 3, 7

[15] Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE MultiMedia*, 24(1):93–96, 2017. 6

[16] Xiaopeng Li, Wenyuan Xu, Song Wang, and Xianshan Qu. Are you lying: Validating the time-location of outdoor images. In *Applied Cryptography and Network Security: 15th International Conference, ACNS 2017, Kanazawa, Japan, July 10-12, 2017, Proceedings 15*, pages 103–123. Springer, 2017. 2, 7

[17] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3, 7, 2

[18] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *Proceedings of the 40th International Conference on Machine Learning*, pages 23498–23515. PMLR, 2023. 2, 3, 5

[19] Radu Paul Mihail, Scott Workman, Zach Bessinger, and Nathan Jacobs. Sky segmentation in the wild: An empirical study. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–6, 2016. 6

[20] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–579, 2018. 3, 6

[21] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal,

Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 7, 2

[22] Rafael Padilha, Tawfiq Salem, Scott Workman, Fernanda A Andaló, Anderson Rocha, and Nathan Jacobs. Content-aware detection of temporal metadata manipulation. *IEEE Transactions on Information Forensics and Security*, 17:1316–1327, 2022. 1, 2, 6, 7

[23] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. In *European Conference on Computer Vision*, pages 196–215. Springer, 2022. 3, 6

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 7

[25] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019. 3

[26] Tawfiq Salem, Scott Workman, and Nathan Jacobs. Learning a dynamic map of visual appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12435–12444, 2020. 2, 6, 7

[27] Tawfiq Salem, Jisoo Hwang, and Rafael Padilha. Timestamp estimation from outdoor scenes. In *Annual ADFSL Conference on Digital Forensics, Security and Law*, number 2. Embry-Riddle Aeronautical University Commons, 2022. 1, 2, 6, 7

[28] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018. 3, 6

[29] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[30] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020. 3

[31] Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio S Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 20020–20029, 2022. 8, 5

[32] Sirnam Swetha, Mamshad Nayeem Rizve, Nina Shvetsova, Hilde Kuehne, and Mubarak Shah. Preserving modality structure improves multi-modal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21993–22003, 2023. 8, 5

[33] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020. 2

[34] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 6

[35] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 3

[36] Tsung-Hung Tsai, Wei-Cih Jhou, Wen-Huang Cheng, Min-Chun Hu, I-Chao Shen, Tekoing Lim, Kai-Lung Hua, Ahmed Ghoneim, M Anwar Hossain, and Shintami C Hidayati. Photo sundial: estimating the time of capture in consumer photos. *Neurocomputing*, 177:529–542, 2016. 2, 7

[37] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 4, 6, 7, 5

[38] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 2621–2630, 2017. 3, 6

[39] L Warszawski, K Frieler, V Huber, F Piontek, O Serdeczny, X Zhang, Q Tang, M Pan, Y Tang, Q Tang, et al. Center for international earth science information network—ciesin—columbia university.(2016). gridded population of the world, version 4 (gpwv4): Population density. palisades. ny: Nasa socioeconomic data and applications center (sedac). doi: 10. 7927/h4np22dq. *Atlas of environmental risks facing China under climate change*, page 228, 2017. 5

[40] Scott Wehrwein, Kavita Bala, and Noah Snavely. Shadow detection and sun direction in photo collections. In *2015 International Conference on 3D Vision*, pages 460–468. IEEE, 2015. 3

[41] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 37–55. Springer, 2016. 3, 6

[42] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2015. Acceptance rate: 30.3%. 6

[43] Angelos Zavras, Dimitrios Michail, Begüm Demir, and Ioannis Papoutsis. Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment. *arXiv preprint arXiv:2402.09816*, 2024. 3

[44] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[45] Menghua Zhai, Tawfiq Salem, Connor Greenwell, Scott Workman, Robert Pless, and Nathan Jacobs. Learning geo-temporal image features. In *British Machine Vision Conference*, 2019. 1, 2, 6, 8, 4

[46] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1

[47] Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2749–2754, 2024. 3, 6

[48] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 5

[49] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 3

[50] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022. 3